Termproject

Introduction

This is a causal analysis on the properties on sale in Budapest. The data was scraped from ingatlan.com by me on the 16th of December, 2021. I investigate the effect of the size of the property on the price of the property.

I think analysing housing prices is always meaningful, as it is a crucial part of our daily life, as everyone needs to live somewhere. People buy sell and rent properties all the time, which is influenced by the prices of the properties. Understanding the change in the prices offers us the opportunity to have further investigations. My goal is to see whether on average if larger properties are more expensive or not then smaller properties.

Data

The data is from ingatlan.com. It was scraped on the 16th of December, 2021. The scrapper went through every property listed under https://ingatlan.com/szukites/elado+lakas+budapest. It scraped, the address, the price, the number of rooms, the description, and the 20 element long list below the description. We have 27908 observations and 27 variables. The scrape and the cleaning are more detailed in the appendix. TODO appendixbe kep es a scraper es clean filter.

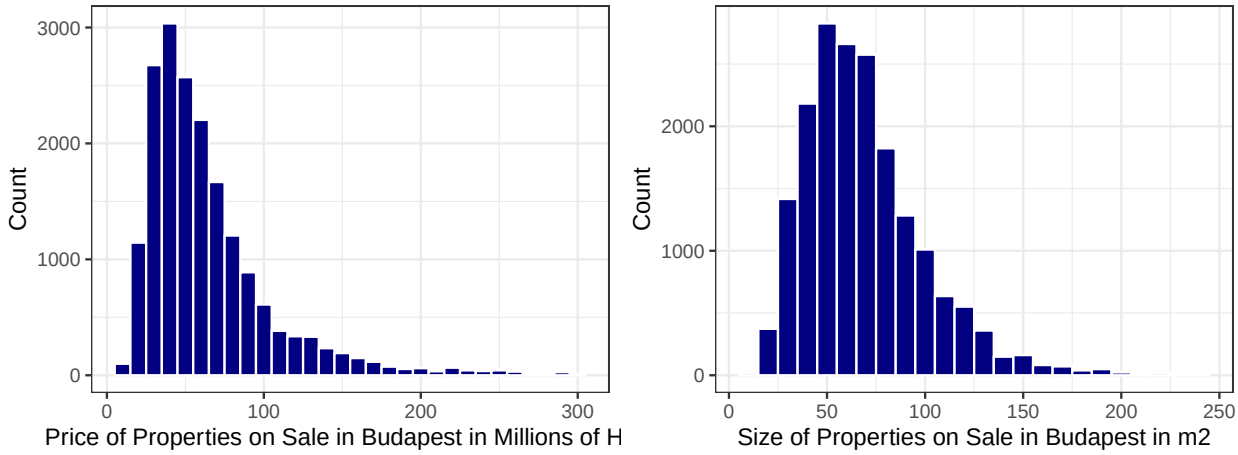
Table 1: Descriptive statistics

	Mean	Median	SD	Min	Max	P05	P95
Price	65.12	54.50	41.68	6.90	299.99	23.50	147.50
Size	69.60	64.00	31.10	11.00	240.00	30.00	127.00
Number of Rooms	2.47	2.00	1.05	1.00	11.00	1.00	4.00
Building Year	1974.52	1980.00	27.03	1950.00	2025.00	1950.00	2022.00
Condition	3.45	3.00	1.36	1.00	6.00	1.00	6.00

The number of observations is 18343 for all of our key variables. These are after filtering on price so that it is greater than 5M HUF, but less then 300M HUF. Also filtered on size so it is less greater than 10 m2, but less than 250 m2. I choose these values as I feel like these are filtering for habitable houses. Usually bigger and more expensive properties were multiple houses, for investing. While smaller and cheaper properties were not really houses, but part of houses and people looking for swap houses.

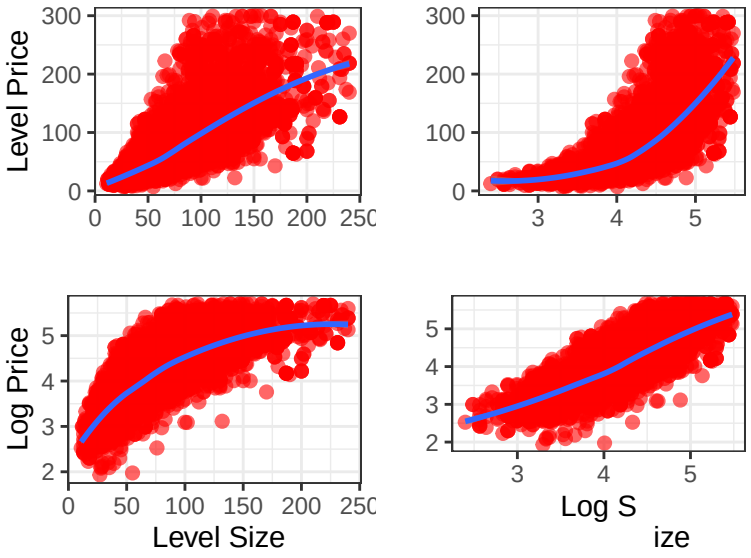
These are the summary stats of the already filtered dataset, so as we can see the minimums and maximum are close to the filter values. The mean of the price is 65.12M HUF the median is 54.5M HUF this means we have a long right tail. The mean of the Size is 69.60 m2 the median is 64.0 m2 this means we have a long right tail.

As the focus are on the price and on the size, the next Figure shows the histograms for these variables. I show the rest of the histograms in the appendix. TODO



We can see the long right tail on the histograms.

We can see the level, and log associations of price and area. The key pattern of associations are:



How will you include this in your model? TODO Short description on the other variables: 2-10 sentence depends on the amount of variables you have. You should reference your decisions on the graphs/analysis which are located in the appendix.

Model

TODO My preferred model is:

$\log(\text{price}) = 5.41 + 0.97\ (log(size) < 4,5)\ 1.18\ (logarea \geq 18) + \delta Z$

where Z are standing for the controls, which includes controlling for year built, number of rooms, and condition. From this model we can infer:

- 5.41 is hard to interpret as both sides are in log, so we are more interested in the differences

- when the size of properties is 10 percent larger, but below the log value of 4.5, we see properties to cost on average 0.97 percent more.
- when the size of properties is 10 percent larger, but the log value is above or equal to 4.5, we see properties to cost on average 1.18 percent more.

Based on the heteroskedastic robust standard errors, these results are statistically different from zero. To show that, I have run a two-sided hypothesis test:

$$H_0 := \beta_1 = 0$$

$$H_A := \beta_1 \neq 0$$

I have the t-statistic as 94.25 and the p-value as 0, which confirms my conclusion.

We compare multiple models to learn about the stability of the parameters:

[H]

Table 2: Models to uncover relation between Size of Properties and Price of Properties

	(1)	(2)	(3)	(4)	(5)
Intercept	-0.3550***	-0.2318***	-6.804***	-6.462***	5.410***
	(0.0225)	(0.0276)	(0.1616)	(0.1757)	(0.3168)
Log Size	1.053***				
	(0.0055)				
Log Size (<4.5)		1.022***	1.021***	0.9625***	0.9709***
		(0.0069)	(0.0068)	(0.0112)	(0.0103)
Log Size (>=4.5)		1.181***	1.203***	1.208***	1.182***
		(0.0219)	(0.0220)	(0.0278)	(0.0251)
Building Year			0.0033***	0.0032***	
			(8.05e-5)	(8.36e-5)	
Number of Rooms < 4				0.0315***	0.0076
				(0.0048)	(0.0044)
Number of Rooms >= 4				-0.0429***	-0.0346***
				(0.0087)	(0.0082)
Building Year <1990					-0.0029***
					(0.0002)
Building Year >=1990					0.0070***
					(0.0003)
Condition					0.0843***
					(0.0022)
Observations	18,343	18,343	18,343	18,343	18,343
R2	0.68353	0.68437	0.71052	0.71161	0.76310

Robustness check / 'Heterogeneity analysis'

Task: calculate and report t-tests for each countries. TODO

Conclusion

TODO HERE COMES WHAT WE HAVE LEARNED AND WHAT WOULD STRENGHTEN AND WEAKEN OUR ANALYSIS.

Appendix

TODO Here comes all the results which are referenced and not essential for understanding the MAIN results.